Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Автоматизированные системы обработки информации и управления»

Отчет

Рубежный контроль № 1

По курсу «Технологии машинного обучения»

ИСПОЛНИТЕЛЬ:

Горбатенко И.А.

Группа ИУ5-64

_____

"__"_____2020 г.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

_____

"__"_____2020 г.

Москва 2020

Рубежный контроль №1 по курсу "Технологии машинного обучения"

Горбатнко И.А. ИУ5-64

Задание:

Задача №1. Для заданного набора данных проведите корреляционный анализ. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

Выполнение:

К сожалению я не разобрался с датасетом 3 варианта, потому что не нашел заголовки атрибутов, а там все на английском, и это не такая простая задача. Поэтому в данном задании будет использован датасет 6 варианта Admission_Predict.csv

Импортируем библиотеки:

In [2]:
```
import os
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsRegressor, KNeighborsClassifier
from sklearn.metrics import accuracy_score, balanced_accuracy_score
from sklearn.metrics import precision_score, recall_score, f1_score, classi
from sklearn.metrics import confusion_matrix
from sklearn.metrics import plot_confusion_matrix
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import mean_absolute_error, mean_squared_error, mean_s
from sklearn.metrics import roc_curve, roc_auc_score
from sklearn.svm import SVC, NuSVC, LinearSVC, OneClassSVM, SVR, NuSVR, Lin
from sklearn.tree import DecisionTreeClassifier, DecisionTreeRegressor, exp
from sklearn.ensemble import RandomForestClassifier, RandomForestRegressor
from sklearn.ensemble import ExtraTreesClassifier, ExtraTreesRegressor
from sklearn.ensemble import GradientBoostingClassifier, GradientBoostingRe
from gmdhpy import gmdh
%matplotlib inline
sns.set(style="ticks")
```

Зададим выборку:

In [3]:
```
our_data = pd.read_csv('Admission_Predict.csv', sep=",")
```

Проверим правильность создания выборки:

In [4]:
```
our_data.head()
```

Out[4]:

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	1	337	118	4	4.5	4.5	9.65	1	0.92
1	2	324	107	4	4.0	4.5	8.87	1	0.76
2	3	316	104	3	3.0	3.5	8.00	1	0.72
3	4	322	110	3	3.5	2.5	8.67	1	0.80
4	5	314	103	2	2.0	3.0	8.21	0	0.65

Проверим типы данных:

In [6]:
```
our_data.dtypes
```

Out[6]:
```
Serial No.             int64
GRE Score              int64
TOEFL Score            int64
University Rating      int64
SOP                  float64
LOR                  float64
CGPA                 float64
Research               int64
Chance of Admit      float64
dtype: object
```

Проверяем датасет на наличие пустых значений:

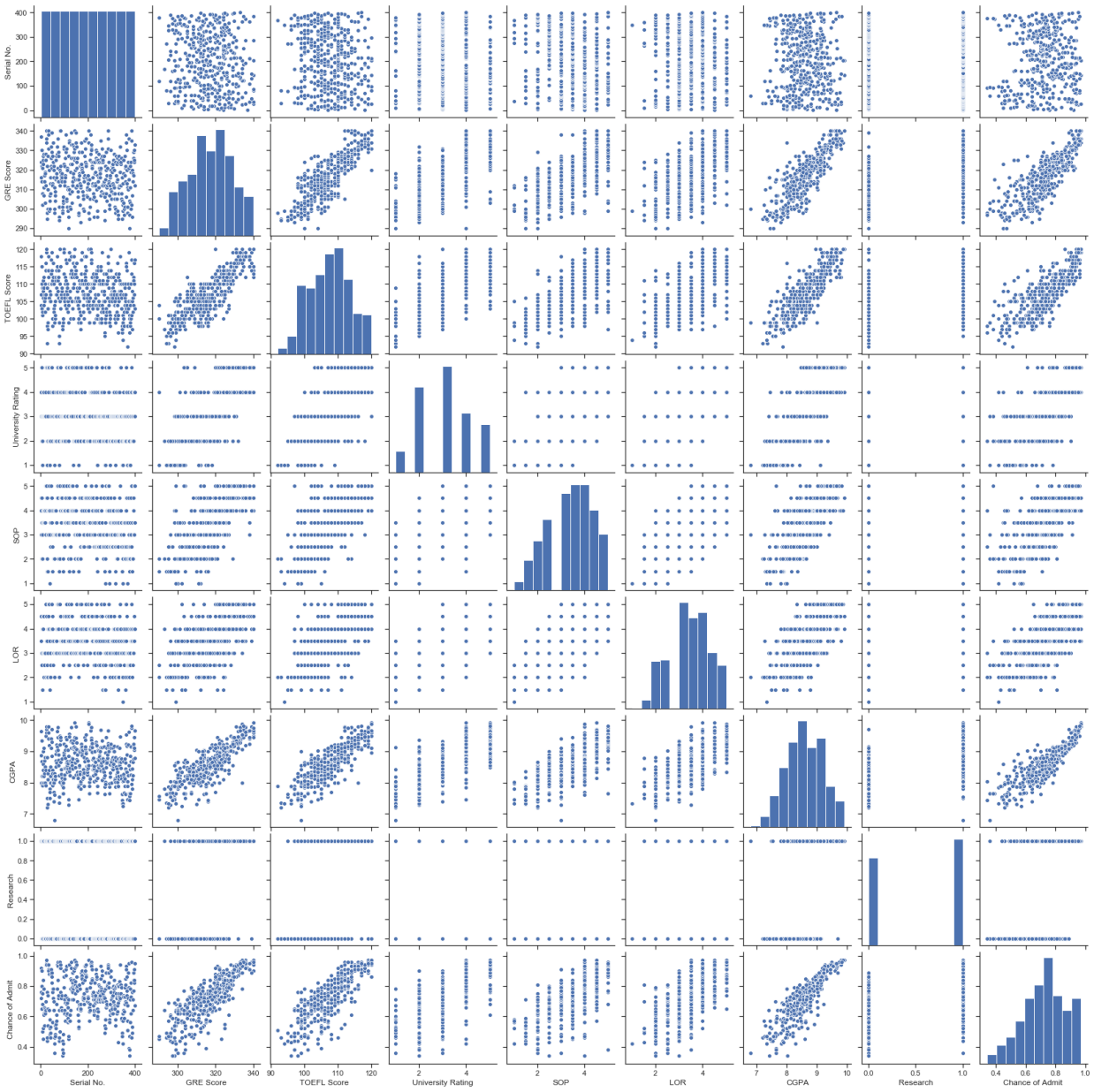
In [7]:
```
our_data.isnull().sum()
```

Out[7]:
```
Serial No.           0
GRE Score            0
TOEFL Score          0
University Rating    0
SOP                  0
LOR                  0
CGPA                 0
Research             0
Chance of Admit      0
dtype: int64
```

Построим парную диаграмму для наглядности структуры наших данных:

In [8]:
```
sns.pairplot(our_data)
```

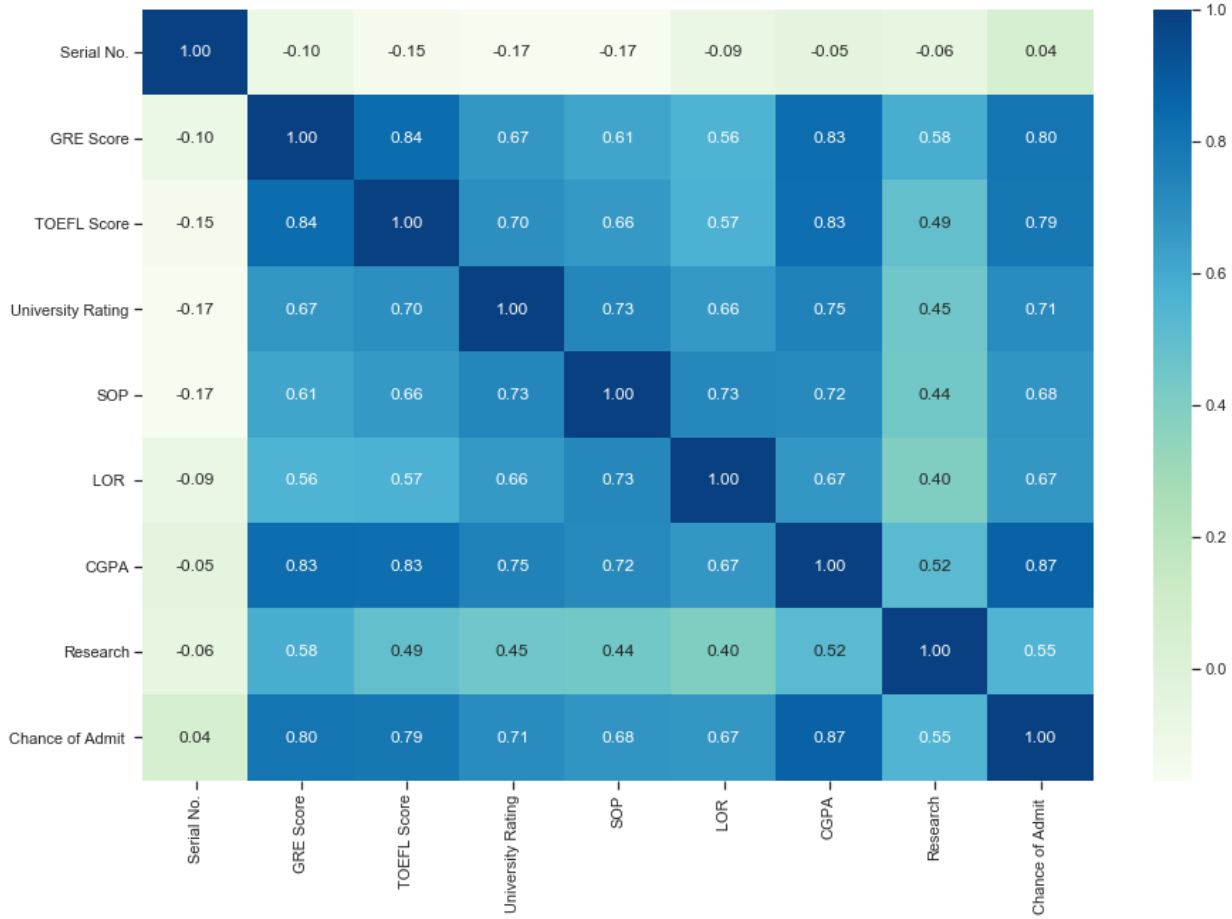
Out[8]:
```
<seaborn.axisgrid.PairGrid at 0x1a173eb410>
```



Построим корреляционную матрицу:

In [10]:
```
fig, ax = plt.subplots(figsize=(15,10))
sns.heatmap(our_data.corr(), annot=True, fmt='.2f', cmap='GnBu')
```

Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x101ac5f90>



Выводы: у некоторых признаков отчетливо видна почти линейная связь. Целевой признак хорошо коррелирует пояти со всеми признаками, за исключением Research и Serial No, которые соответственно не стоит включать в модель. CGPA нужно будет включить в модель, но он сильно коррелирует с признаками. Я думаю, что модель классификации построить можно, но нужно попробовать различные комбинации параметров, поскольоку почти все параметры, коррелирующие с целевым признаком, достаточно сильно коррелируют между собой, поэтому вполне вероятно, что наиболее точной моделью окажется даже та, в которую включен всего лишь один признак - CGPA